인물에 독립적인 표정인식을 위한 Action Unit 기반 얼굴특징에 관한 연구

이승호, 김형일, 박성영, 노용만
한국과학기술원 전기및전자공학과
e-mail : {leesh09, hyungil.kim, clover200, ymro}@kaist.ac.kr

Action Unit Based Facial Features for Subject-independent Facial Expression Recognition

Seung Ho Lee, Hyung-Il Kim, Sung Yeong Park, Yong Man Ro
Dept. of Electrical Engineering (EE), Korea Advanced Institute of Science and Technology (KAIST)

요 약

실제적인 표정인식 응용에서는 테스트 시 등장하는 인물이 트레이닝 데이터에 존재하지 않는 경우가 빈번하여 성능 저하가 발생한다. 본 논문에서는 인물에 독립적인(subject-independent) 표정인식을 위한 얼굴특징을 제안한다. 제안방법은 인물에 공통적인 얼굴 근육 움직임(Action Unit(AU))에 기반한 기하학 정보를 표정 특징으로 사용한다. 따라서 인물의 고유 아이덴티티(identity)의 영향은 감소되고 표정과 관련된 정보는 강조된다. 인물에 독립적인 표정인식 실험결과, 86%의 높은 표정인식률과 테스트 비디오 시퀀스 당 3.5ms(Matlab 기준)의 매우 빠른 분류속도를 달성하였다.

1. 서론

얼굴 표정은 사람의 감정(emotion)을 전달하는 중요 매개체로서 이를 자동으로 인식하는 표정인식 기술이 지능형 강의시스템, 자동 광고나 마케팅 등 다양한 응용들에서 활용되고 있다.

표정인식에서 가장 중요한 단계 중 하나는 얼굴특징 추출이다. 최근에 제안된 대부분의 표정인식 방법들은 얼굴의 윤곽이나 주름 등의 텍스처(texture) 정보를 특징으로 사용한다. 대표적으로 local binary pattern(LBP)[1]과 local phase quantization(LPQ)[2]가 있다. 비디오 시퀀스에 내재된 얼굴의 다이나믹(dynamic) 정보를 인식에 사용하기 위해 LBP-TOP(three orthogonal planes)[3]와 LPQ-TOP[4]가 제안되었고 기존의 LBP 나 LPQ 에 비해 높은 표정인식 성능을 보였다. 이 방법은 2D 기반의 텍스처 추출 연산(LBP 또는 LPQ)을 세 개의 직교하는 평면(XY: 공간적 평면, XT 와 YT: 시간적 평면)에 적용하여 얻은 특징들을 연결(concatenate)하여 시공간 얼굴특징으로 사용하였다[4][5]. 그런데 위에 언급한 네 가지 방법들[1]-[4]은 테스트 시 등장하는 인물이 트레이닝 데이터에 존재하지 않는 경우 얼굴에 포함된 인물의 아이덴티티(identity)와 표정간 혼동에 의해 성능저하가 발생하는 문제점을 가진다.

본 논문에서는 본 논문에서는 인물에 독립적인(subject-independent) 표정인식을 위한 새로운 얼굴특징을 제안한다. 제안방법의 특장점은 다음과 같이 두 가지로 요약된다.

1) 제안방법은 사람이 특정 표정(예 : Surprise)을 지을 때 공통적으로 나타나는 얼굴 근육 움직임(Action Unit(AU)[5])에 기반한 기하학 정보를 표정 특징으로 사용한다. 따라서 인물의 고유 아이덴티티(identity)의 영향은 감소되고 표정과 관련된 정보는 강조되어 표정 분류 시 분별력을 높인다.
2) 얼굴특징 추출 시 29 종류의 기하학 특징만을 사용하므로 특징벡터의 차원수를 대폭 감소시킬 수 있다. 표정 데이터를 적은 저장공간으로도 보존할 수 있기 때문에 방대한 양의 표정 데이터를 처리 및 저장해야 하는 대용량 감성인식이나 생체인식 시스템에 적용이 용이하다.

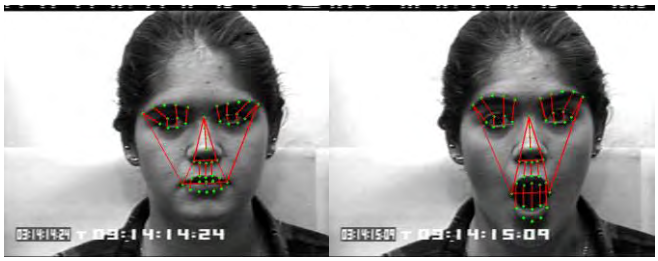
대표적인 공용 데이터베이스인 Cohn-Kanade plus(CK+)를 이용하여 인물에 독립적인 인식조건에서 실험을 수행한 결과, 제안방법은 대표적인 표정특징인 LBP-TOP 와 LPQ-TOP 보다 높은 표정인식률을 달성하였다.

본 논문의 구성은 다음과 같다. 2 장에서는 제안하는 AU 기반 얼굴특징 추출 및 표정인식 방법에 대해 자세히 설명한다. 3 장에서는 실험결과 및 분석을 보이고 4 장에서 결론을 맺는다.

2. 제안하는 얼굴특징 추출 및 표정인식 방법

2.1. 제안하는 Action Unit 기반 얼굴특징 추출

본 절에서는 주어진 (트레이닝 혹은 테스트 용) 비디오 시퀀스에서 제안하는 AU 기반 얼굴특



(a) (b)

(그림 1) 검출된 얼굴 랜드마크(점으로 표시)와 랜드마크 간 정의된 거리(선으로 표시) (a)중립(neutral) 얼굴, (b)피크(peak) 표정얼굴

징 추출방법을 설명한다. 제안방법은 인물에 독립적인 표정인식을 위해 인식을 수행할 비디오 시퀀스에서 표정세기(intensity)가 가장 큰 피크(peak) 표정얼굴과 동일한 인물의 중립(neutral)얼굴 간의 차이(difference)를 이용한다. 또한 제안방법은 표정얼굴을 AU 관점에서 분석하기 위해 얼굴의 랜드마크(landmark) 검출을 사용한다. 본 논문에서는 얼굴 랜드마크 검출을 위해 [6]의 방법을 사용하였다. 이 방법을 중립 얼굴과 피크 표정얼굴에 적용하여 얻은 49개의 얼굴 랜드마크 검출결과 예를 그림 1(a)와 (b)에 각각 나타내었다. 다음으로 표정을 지을 때 발생하는 얼굴형태의 변화를 직관적으로 표현하기 위해 29 쌍의 얼굴 랜드마크 간의 거리를 그림 1 과 같이 정의하였다.

중립 얼굴과 피크 표정얼굴 사이에서 발생하는 29 쌍의 얼굴 랜드마크 간 거리의 변화를 이용한 얼굴특징 추출과정을 다음과 같이 설명한다. 먼저 비디오 시퀀스의 피크 표정얼굴에서 i 번째와 j 번째 얼굴 랜드마크 간의 거리 $(dist_{i,j}^{\text{peak}})$를 수식 (1)과 같이 정의한다.

$$dist_{i,j}^{\text{peak}} = \frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{\sqrt{(x_{\text{eyeL}} - x_{\text{eyeR}})^2 + (y_{\text{eyeL}} - y_{\text{eyeR}})^2}}, \quad (1)$$

여기서 x_i와 y_i는 각각 i 번째 얼굴 랜드마크의 수평, 수직 좌표이다. 그리고 x_{eyeL}, x_{eyeR}, y_{eyeL}, y_{eyeR}는 각각 왼쪽과 오른쪽 눈의 수평 좌표, 왼쪽과 오른쪽 눈의 수직 좌표이다. 양쪽 눈의 좌표는 눈 부위를 둘러싼 6 개의 좌표에 평균을 취하여 얻었다. 수식 (1)에서 분모항은 카메라와 얼굴간의 거리에 따른 얼굴 스케일의 변화를 보상해 주기 위해 삽입되었다.

중립 얼굴과 피크 표정얼굴 사이에서 발생한 i 번째와 j 번째 얼굴 랜드마크 간의 거리변화 $(f_{i,j})$는 수식 (2)와 같이 정의한다.

$$f_{i,j} = dist_{i,j}^{\text{peak}} - dist_{i,j}^{\text{neutral}}, \quad (2)$$

여기서 $dist_{i,j}^{\text{neutral}}$는 중립 얼굴의 i 번째와 j 번째 얼굴 랜드마크 간 거리이며, 피크 표정얼굴과 동일하게 수식 (1)을 이용하여 계산할 수 있다.

수식 (2)에서 얻은 29 개의 얼굴 랜드마크 간 거리 변화 값 $f_{i,j}$을 연결(concatenate)하여 29 차원의 제안하는 AU 기반 얼굴특징 $\mathbf{f} \in \Re^{29\times 1}$를 얻는다.

2.2. 제안하는 얼굴특징을 이용한 sparse representation 분류 기반 표정인식

본 절에서는 최근 얼굴 표정인식을 포함한 패턴인식에서 매우 높은 효용성 입증된 sparse representation(SR) 분류(classification)방법[7]과 2.1 절에서 설명한 방법으로 얻은 얼굴특징을 이용한 표정인식 과정을 설명한다. 먼저 테스트 비디오 시퀀스에서 얻어진 얼굴특징을 $\mathbf{f}^{\text{test}} \in \Re^{29\times 1}$라고 표기한다. 그리고 n 번째 표정 클래스 $(n = 1,...,N)$의 m 번째 $(m = 1,...,M_n)$ 트레이닝 비디오 시퀀스에서 얻어진 얼굴특징을 $\mathbf{f}_{n,m}^{\text{train}} \in \Re^{29\times 1}$이라고 표기한다. SR 분류를 통해 $\mathbf{f}^{\text{test}}$에 대한 표정인식을 수행하기 위해 딕셔너리[7]를 정의한다. 딕셔너리는 수식 (3)과 같이 정의된다.

$$\mathbf{A} = [\mathbf{f}_{1,1}^{\text{train}},...,\mathbf{f}_{1,M_1}^{\text{train}},\mathbf{f}_{2,1}^{\text{train}},...,\mathbf{f}_{2,M_2}^{\text{train}},...,\mathbf{f}_{N,M_N}^{\text{train}}]. \quad (3)$$

$\mathbf{f}^{\text{test}}$에 대한 SR 은 수식 (4)와 같은 ℓ_1-놈 최소화(ℓ_1-norm minimization) 문제를 풀어줌으로써 얻을 수 있다[7].

$$\hat{\mathbf{w}} = \arg\min \|\mathbf{w}\|_1, \quad s.t. \left\|\mathbf{f}^{\text{test}} - \mathbf{A}\mathbf{w}\right\|_2 \le \varepsilon, \quad (4)$$

여기서 ε 은 에러 텀(error term)이다. 본 논문에서는 수식 (4)의 sparse 해(solution) 벡터[7] $\hat{\mathbf{w}}$를 구하기 위해 regularized orthogonal matching pursuit(ROMP)[8]를 사용하였다

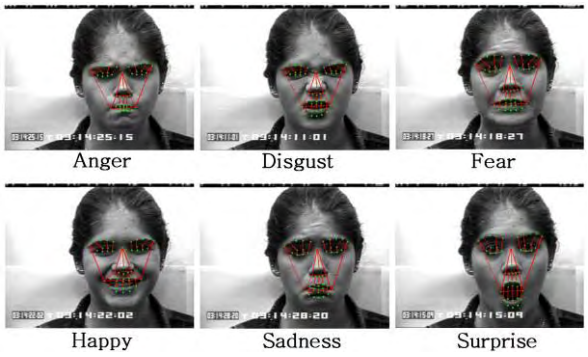
마지막으로 표정 클래스 라벨 n^*은 수식 (5)와 같이 평균 sparse 계수(coefficient)의 크기가 가장 큰 표정 클래스를 찾음으로써 결정된다.

$$n^* = \arg\max_{n=1}^{N} \frac{1}{M_n} \sum_{m=1}^{M_n} w_{n,m}, \quad (5)$$

여기서 $w_{n,m}$는 sparse 해 벡터 $\hat{\mathbf{w}}$에서 n 번째 표정 클래스의 m 번째 트레이닝 비디오 시퀀스에서 추출된 얼굴특징을 나타낸다.

3. 실험결과

제안하는 방법을 평가하기 위해 Cohn-Kanade plus (CK+) 데이터베이스[9]를 사용하였다. 일곱 가지의 감정 중 하나로 라벨링 된 325 개의 비디오 시퀀스 (Anger 45 개, Contempt 18 개, Disgust 58 개, Fear 25 개, Happy 69 개, Sadness 28 개, Surprise 82 개)를 선택하였다. 각각의 비디오 시퀀스는 최소 5 개에서 최대 67 개의



(그림 2) 6 가지 기본 감정(basic emotion)에 대한 비디오 시퀀스의 피크(peak) 표정얼굴 예제

<표 1> Sparse representation 분류 프레임워크에서 제안하는 특징과 대표적인 두 가지 특징들과의 성능 비교

표정인식 방법	표정인식률(%)
LBP-TOP [3] + SRC	78.4
LPQ-TOP [4] + SRC	82.1
제안방법 + SRC	**86.0**

<표 2> 제안방법에 대한 혼동 행렬(confusion matrix). 'Actual'과 'Predicted'는 각각 실제의 표정라벨과 SR 분류를 통해 얻어진 표정라벨을 의미한다

Actual / Predicted	Anger	Contempt	Disgust	Fear	Happy	Sad	Surprise
Anger	**60.0**	5.6	3.5	0.0	0.0	7.1	0.0
Contempt	13.3	**72.2**	1.7	4.0	1.4	7.1	0.0
Disgust	11.1	0.0	**91.4**	0.0	0.0	7.1	0.0
Fear	2.2	5.6	0.0	**72.0**	0.0	14.3	0.0
Happy	2.2	11.1	1.7	16.0	**98.6**	3.6	0.0
Sad	11.1	5.6	1.7	4.0	0.0	**60.7**	0.0
Surprise	0.0	0.0	0.0	4.0	0.0	0.0	**100.0**

프레임을 포함하며 첫 프레임에서 중립(neutral)으로 시작하여 점차 표정의 세기가 강해져서 마지막 프레임에서 피크(peak) 표정을 나타낸다. 본 논문에서는 제안하는 방법에서 각각 비디오 시퀀스의 첫 프레임과 마지막 프레임을 중립 프레임과 피크 프레임으로 사용하였다. 그림 2 는 6 가지 기본 감정에 대한 비디오 시퀀스의 피크 표정얼굴과 29 쌍의 랜드마크 간 거리를 표시한 예를 보여준다. 인물에 독립적인 표정인식 성능 측정을 위해 [10]에서 제시한 방법과 유사하게 leave-one-subject-out(LOSO) 크로스 밸리데이션(cross validation) 방식[10]을 채택하였다. 성능 척도로서 표정인식률(recognition rate)를 사용하였다.

제안하는 표정인식 특징의 효용성을 검증하기 위해 표정인식에서 대표적인 시공간 얼굴특징인 LBP-TOP[3]와 LPQ-TOP[4]가 비교 목적으로 사용되었다. 공정한 비교를 위해 LBP-TOP 와 LPQ-TOP 의 분류는 제안방법과 동일하게 SR 분류방법을 사용하였다.

SR 분류 프레임워크에서 제안방법과 두 가지 특징들과의 표정인식률 비교를 표 1 에 나타냈다. 제안방법은 비교방법에 비해 약 4~6% 높은 표정인식률을 보였다. 시공간의 텍스처에 기반하는 LBP-TOP 와 LPQ-TOP 에 비해 제안방법은 중립표정 대비 변화하는 AU 기반 특징을 추출한다. 이 때문에 인물에 독립적인 순수 표정정보를 인식에 사용하여 성능 향상을 달성할 수 있었다. 제안방법 성능에 대한 혼동 매트릭스(confusion matrix)를 표 2 에 나타내었다.

4. 결론

본 논문에서는 인물에 독립적인(subject-independent) 표정인식을 위한 얼굴특징을 제안하였다. 제안방법은 테스트에 등장하는 인물에 관계 없이 표정과 관련된 성분을 추출하기 위해 Action Unit(AU)에 기반한 기하학 특징을 정의하였다. 인물에 독립적인 표정인식에서 실험을 수행한 결과, 86%의 높은 표정인식률을 달성하였으며 특징의 적은 차원 수에 의해 Matlab 환경에서 테스트 비디오 시퀀스 당 평균 3.5ms 의 매우 빠른 분류 속도를 보였다.

감사의 글

이 논문은 2014 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2011-0011383).

참고문헌

[1] T. Ahonen, A. Hadid, and M. "Face Description with Local Binary Pattern: Application to Face Recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 12, pp. 2037-2041, 2006.

[2] C. H. Chan, J. Kittler, N. Poh, T. Ahonen, and M. "(Multiscale) Local Phase Quantization Histogram Discriminant Analysis with Score Normalisation for Robust Face Recognition," IEEE Int'l Conf. on Computer Vision Workshop, 2009.

[3] G. Zhao and M. Pietikäinen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 6, 2007.

[4] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, "A Dynamic Appearance Descriptor Approach to Facial Actions Temporal Modeling," IEEE Trans. Syst., Man, Cybern, B, vol. 44, no. 2, pp. 161-174, 2014.

[5] Y. L. Tian, T. Kanade, and J. F. Cohn, "Facial Expression Analysis," Handbook of Face Recognition, S. Z. Li, and A. K. Jain, eds., pp. 247-276, Springer, 2011.

[6] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental Face Alignment in the Wild," IEEE Int'l Conf. on Computer Vision and Pattern Recognition, 2014.

[7] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust Face Recognition via Sparse Representation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 2, pp. 210-227, 2009.

[8] D. Needell, and R. Vershynin, "Uniform Uncertainty Principle and Signal Recovery via Regularized Orthogonal Matching Pursuit," Foundations of Computational Mathmatics, vol. 9, no. 3, 2009.

[9] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, and Z. Ambadar, "The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression," IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR), 2010.

[10] S. H. Lee, K. N. Plataniotis, and Y. M. Ro, "Intra-Class Variation Reduction Using Training Expression Images for Sparse Representation Based Facial Expression Recognition," IEEE Trans. Affective Computing, vol. 5, no. 3, pp. 340-351, 2014.